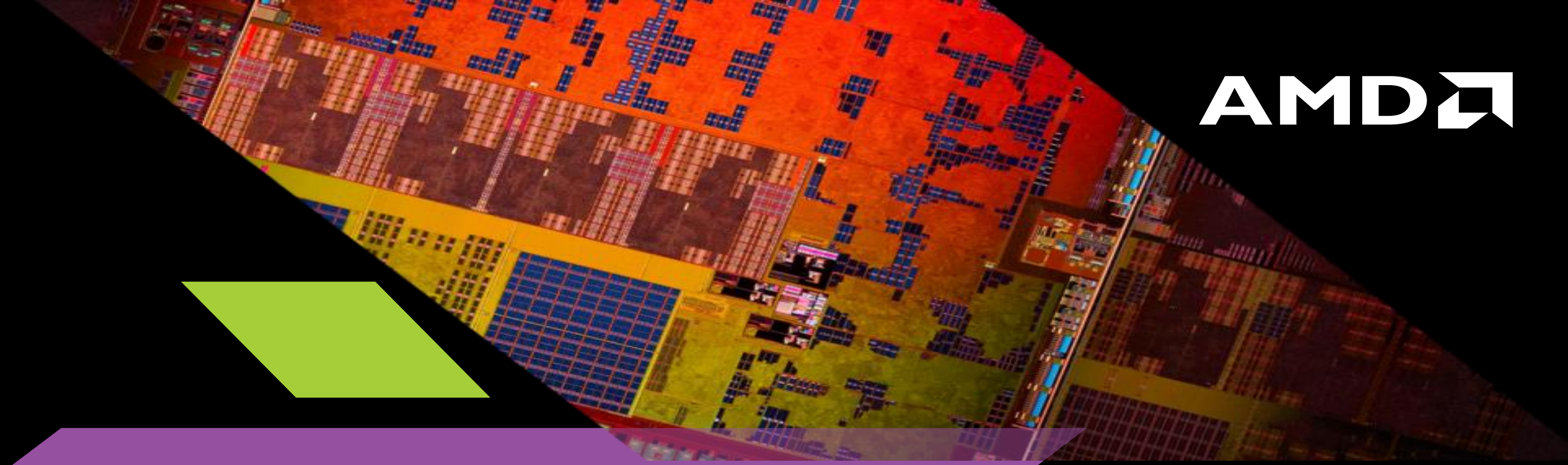


MEMORY TECHNOLOGY AND APPLICATIONS

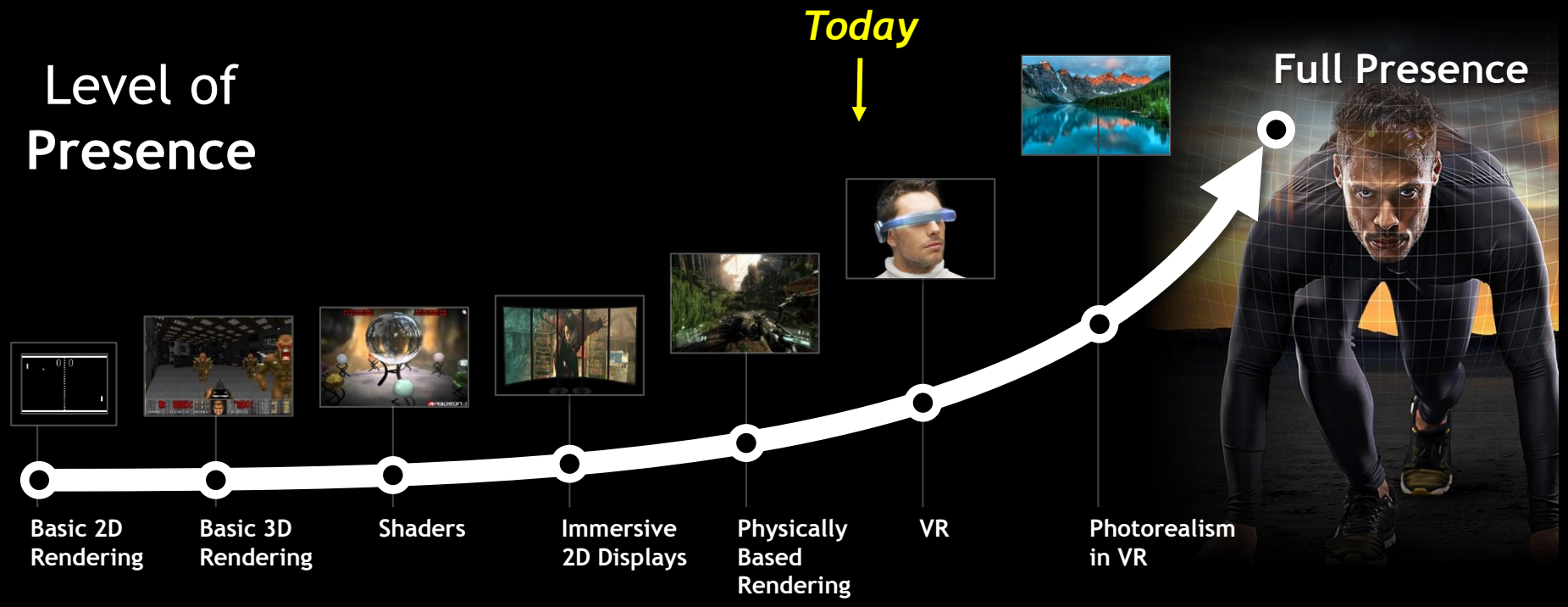
ALLEN RUSH
FELLOW, AMD

AGENDA

- Introduction
- Memory Resources in High Performance Compute Applications
- Challenges in Emerging Applications
- Future possibilities for memory-PE arrangements

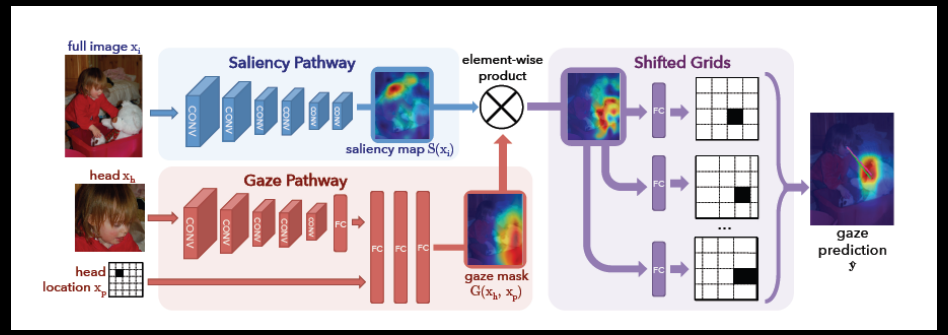
MEMORY REQUIREMENTS IN EMERGING APPLICATIONS

- Example: VR
  - Frame rate, resolution, FOV, DR all contributing to higher capacity and BW

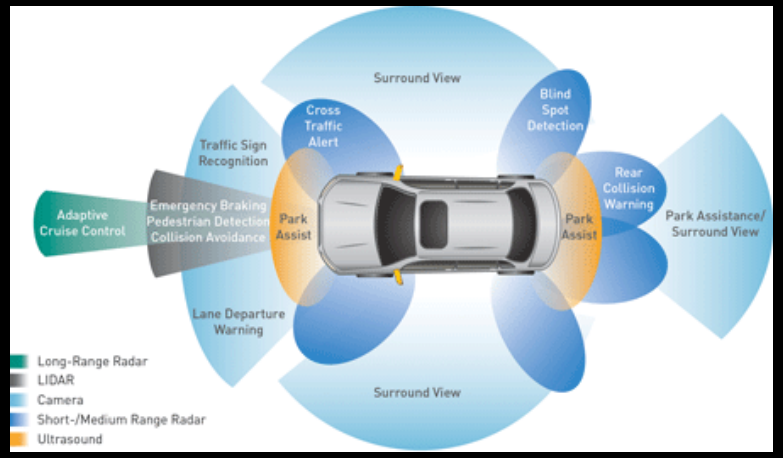


AUTOMOTIVE: INTELLIGENT VEHICLES ARE A REALITY

- Throughput from sensors, recall from trained databases, big data updates…

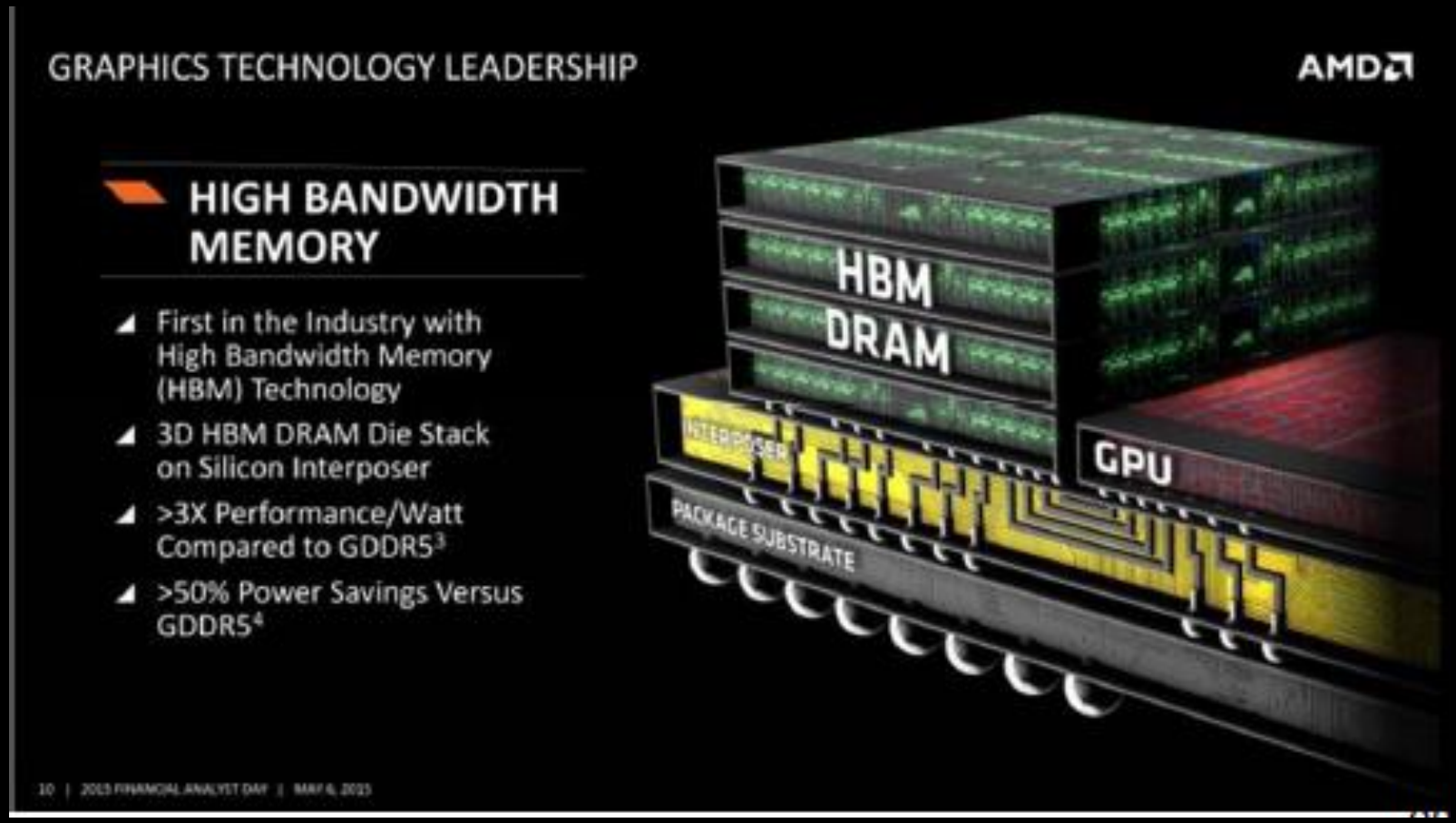


Gaze detection, learning and prediction

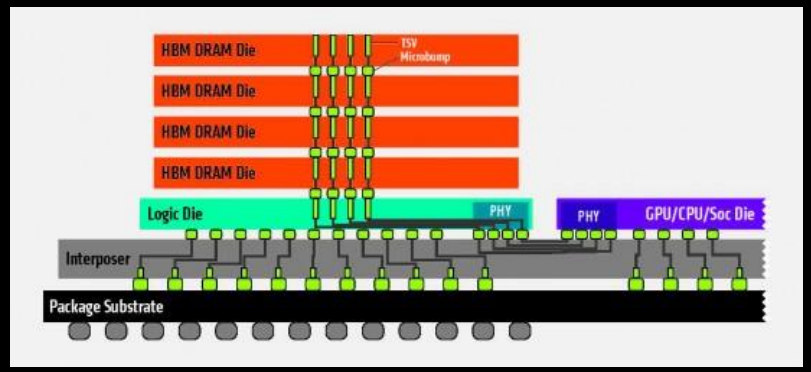


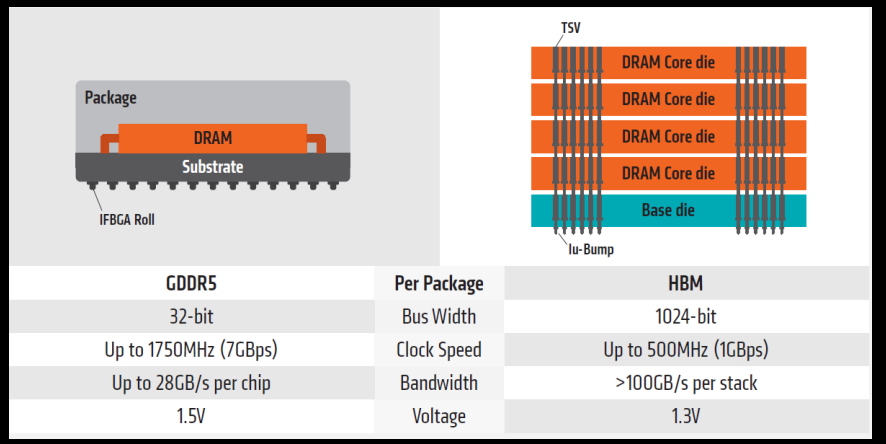
Automotive safety: obstacles, cues, dictionary detection

PACKAGING AND MEMORY ARCHITECTURES TO MATCH HIGH DENSITY COMPUTE



HBM INTRODUCTION – SIZE, BW, POWER, INTERCONNECT: IMPROVEMENTS FOR APPLICATIONS ACROSS THE BOARD





GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V

AMD Radeon™ R9 Nano graphics card. Small size. Giant impact.

- The world's first small form factor (6-inch) graphics card with High-Bandwidth Memory (HBM) delivering new advances in power efficiency.
- Powerful performance for unbelievably "real" 4K and VR gaming.
- A new paradigm for the Mini-ITX PC.

REAL TIME VISION AND AUDIO PRESENCE

PHOTOREALISTIC AND FULL 3D ACOUSTICS FOR GAMING, VR, AR

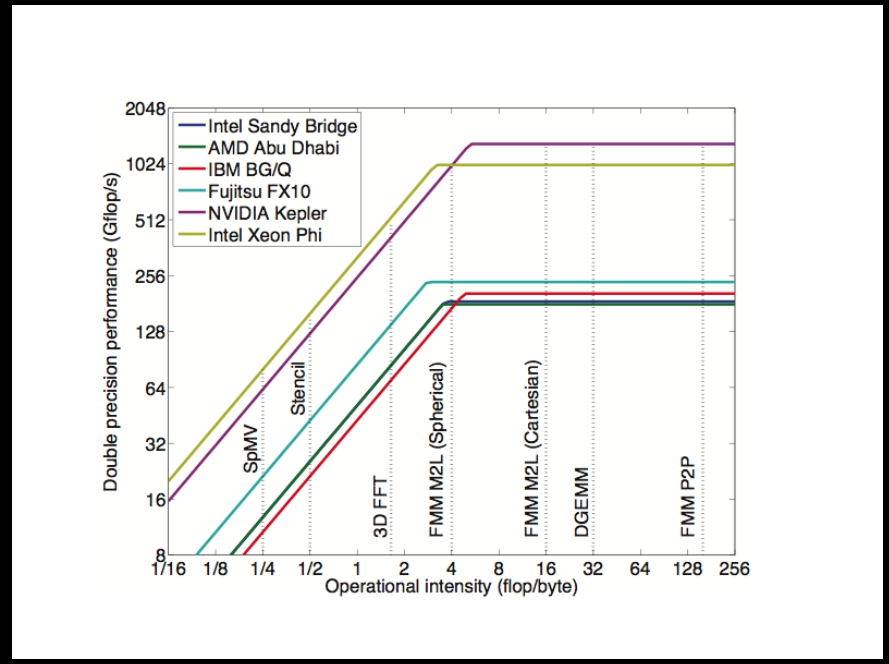
Sensory Integration

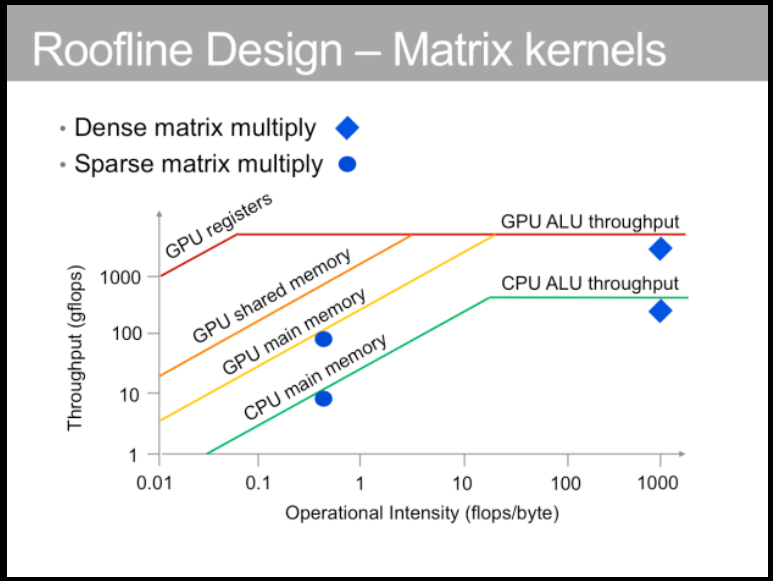
To achieve the vision of full presence you need

Scalable CPUs, GPUs, Accelerators, High BW Memory

ROOFLINE PERFORMANCE MODEL
MATCHING BW TO PEAK COMPUTE PERFORMANCE

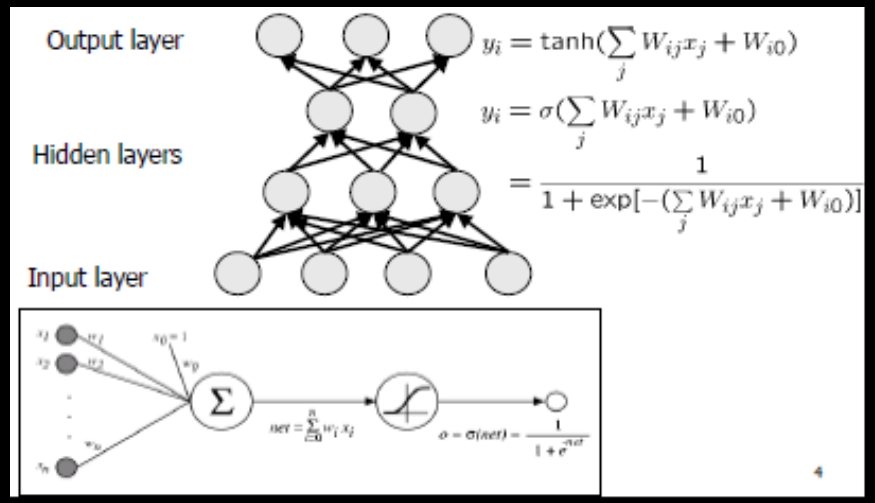
- Peak performance depends on application memory density and BW
- Arithmetic intensity – ratio of actual ops (add, mul…) to mem ops (loads, stores)
- Applications with high AI: maximize peak performance

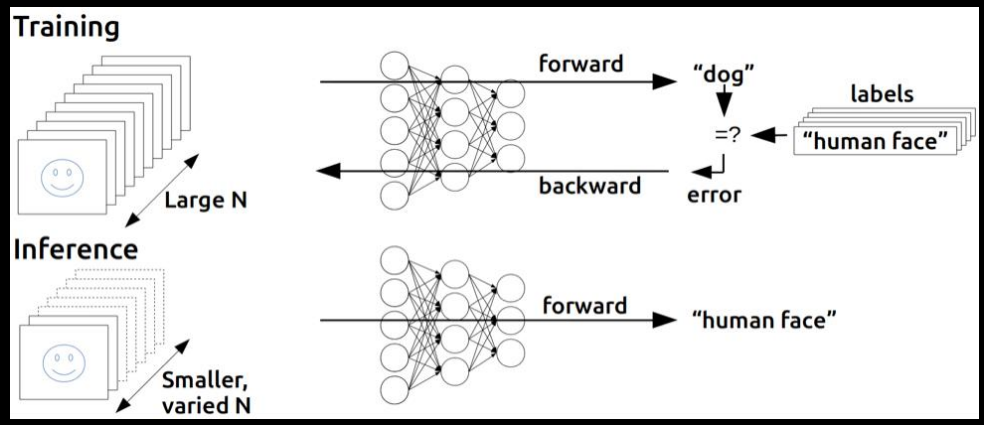




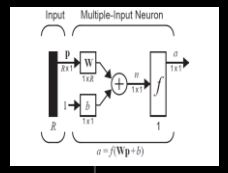
DEEP LEARNING – EMERGING ALGORITHMS FOR ML

- DNN rapidly becoming reliable solution for Machine Learning
- CNN and derivatives are contemporary forms that produce accurate results for many classification problems
- Memory BW and capacity
  - Different for Forward (Inference) and Backward (training)
  - For Training: 1M+ image training data base; 100-1000 parameters to train
  - For Inference: 4x 30fps 4K video ->~1B pixels/s; parameters fixed, variable data

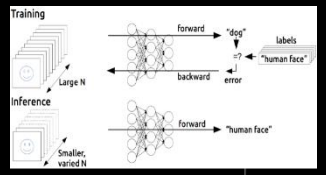




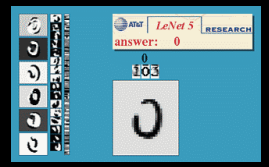
ML PROGRESSION



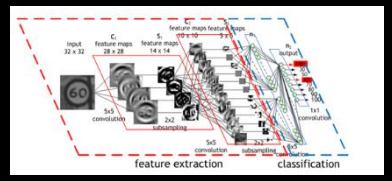
Early ANN
Simple NN
models



Backward
Propagation
Supervised learning



Handwriting
recognition
Deep Belief
Networks



CNN and Deep NN
Accuracy
Image Classification
"**Train to Learn**"

Present
Day

Complex (very)
deep
NN, combinations
of RNN, RCNN,
LSTM…
Real time
Inference
Fast Training

Real-time
scene interp
and residual
learning
"**Learn to
Train**"

DEEP LEARNING APPLICATIONS
AND MANY MORE

Immersive Gaming:
Learned Behavior

Medicine: Learning to
interpret scans

Log in Authentication:
Learning pose and shape

Vision Classification
and Recall

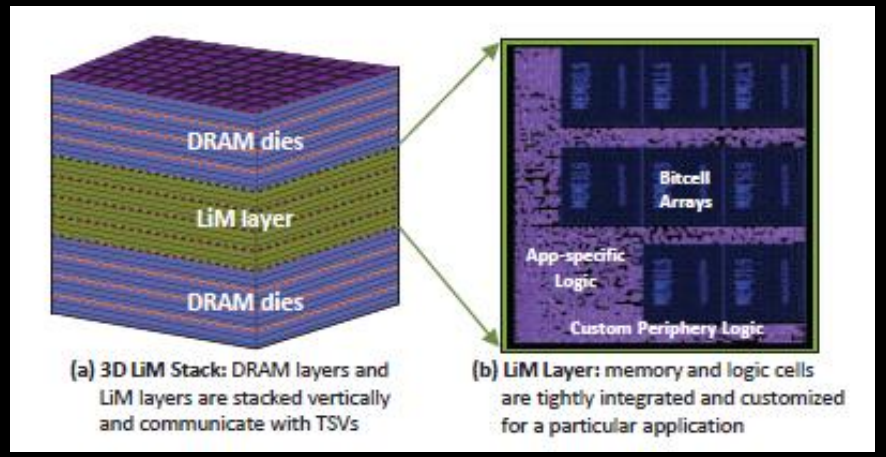
Security

Advanced Driver
Assistance: "Thinking Cars"

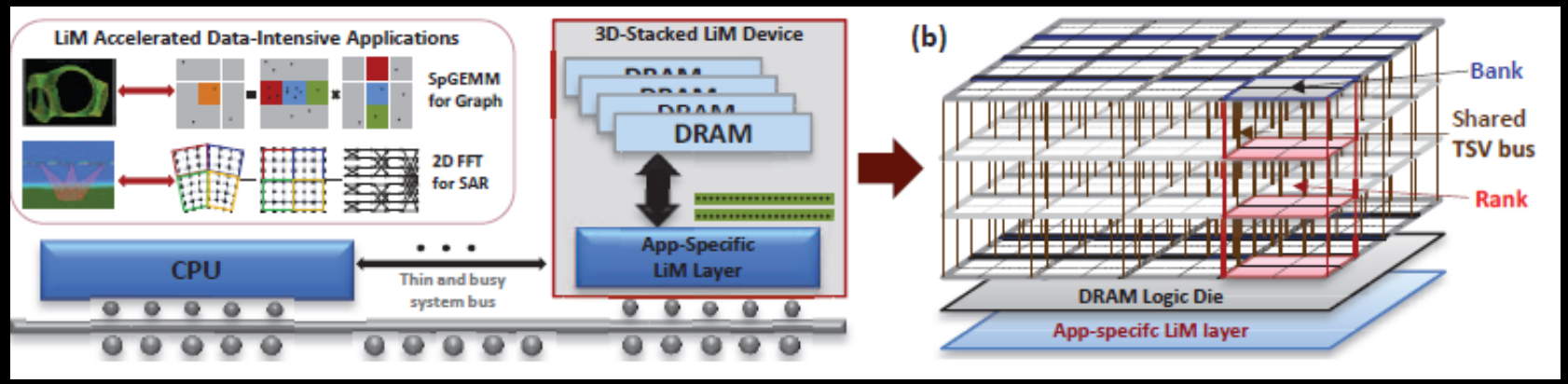
FUTURE REQUIREMENTS

CREATIVE WAYS TO CONSTRUCT DENSE COMPUTE-MEMORY MODULES

- Hard to tell: BW, capacity, power advances – limit?
- Need to get PEs close to L3 or PM
- Novel approaches: LIM

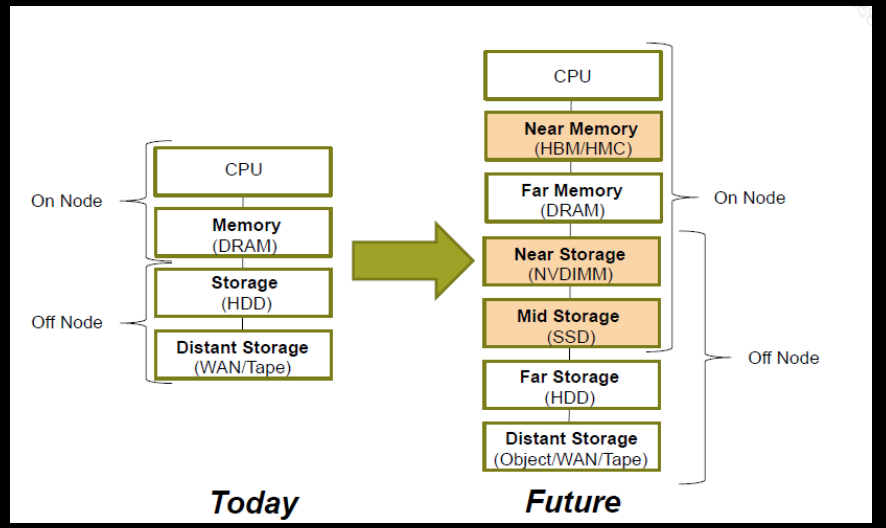


Logic-in-Memory



BIG DATA REQUIREMENTS

- Increasing size and demand for data access-speed of retrieval and storage
- DL NN structures with ~1B parameters and >100M input frames
- Sort, match and classify in near real time
- New training or inference data uploaded at ~2000GB/s



SUMMARY

AMD

- Factors driving advanced memory designs
  - Applications requiring real time video, VR, advanced graphics
- Increased CPU/GPU performance
  - Need for balancing BW, capacity
  - HBM solutions
- Emerging applications with unique memory requirements
  - ML – training and inference
- Novel solutions for PE-Mem structures
- Big Data
  - More data upload
  - Cloud DL: massive parameter and training data sets